

**CONFERENCE OF EUROPEAN STATISTICIANS**

**UNECE Work Session on Statistical Data Editing**  
(27 – 29 May 2002, Helsinki, Finland)

Topic (ii): Measuring and evaluating data editing quality

**GENERAL PRINCIPLES FOR DATA EDITING IN BUSINESS SURVEYS  
AND HOW TO OPTIMIZE IT**

**Contributed paper**

Submitted by the University of Southampton, United Kingdom<sup>1</sup>

**Abstract:** The aim of this paper is to analyse the problem of data editing in general, taking into account many dimensions, particularly in business surveys, and to propose an overall methodology to optimize data editing. We begin with giving an overall description of the data editing process: main concepts, main stages, etc. Then we focus on the contents of the automatic editing program, that selects the doubtful returns, and we show that there are different levels of acceptability of data. We elaborate on the concept of prioritization of manual checking, and we try to determine what could be a stopping criterion in the process. Then we explain what are the main characteristics of the work of survey operators as far as data editing is concerned, and how to use data editing indicators in order to improve the next survey process. In the last sections we try to describe how the whole process could be optimized, and what are the practical problems in the case of complex business surveys.

**I. INTRODUCTION**

1. In the business survey process, data editing comes after data collection and data keying. As defined by UNECE (United Nations 1997), data editing is “the activity aimed at detecting and correcting errors in data”. Error detection is performed in two ways: automatically (error detection program) or manually (clerical work). Error correction is also performed either manually or automatically (imputation).
2. As there is a lot of human intervention on data, data editing is probably the most expensive part of the survey process: in business surveys, it is estimated at 40% of the total cost of the survey, as indicated by several authors. It turns out that editing has often been costly, and that the costs were not reduced significantly through time, particularly in complex business surveys.
3. On the other hand the literature is unanimous to point out that most of the time, a majority of the checks are inefficient: a minority of records are actually changed. In business surveys, it also happens that even among those erroneous, few changes account for the major part of the total change. Hence it should be possible, at least in theory, to reduce the costs substantially.
4. As far as data editing is concerned, there is a sort of trade-off between quality and cost. Intuitively, if we check a lot of questionnaires, it is costly but it improves quality. If there is less editing, the survey is less expensive, better in terms of timeliness, but the risk is clearly to leave unchanged erroneous data that are influential.

---

<sup>1</sup> Prepared by Pascal Rivière (riviere@socsci.soton.ac.uk).

5. According to different studies, this is not necessarily true: data editing does not always improve data. In business surveys, it is sometimes impossible to improve data, simply because the respondent does not know the true value: this is often due to the inaccuracy of the accounting system of the establishment.

6. Therefore, there seems to be a discrepancy between two facts: on one hand, not only is data checking often inefficient, but data correction is also very imperfect; it means that we should lower dramatically the amount of editing. On the other hand, even if there are some success stories, few improvements were observed in terms of reduction of editing costs in business surveys, particularly in large surveys (many units, many variables).

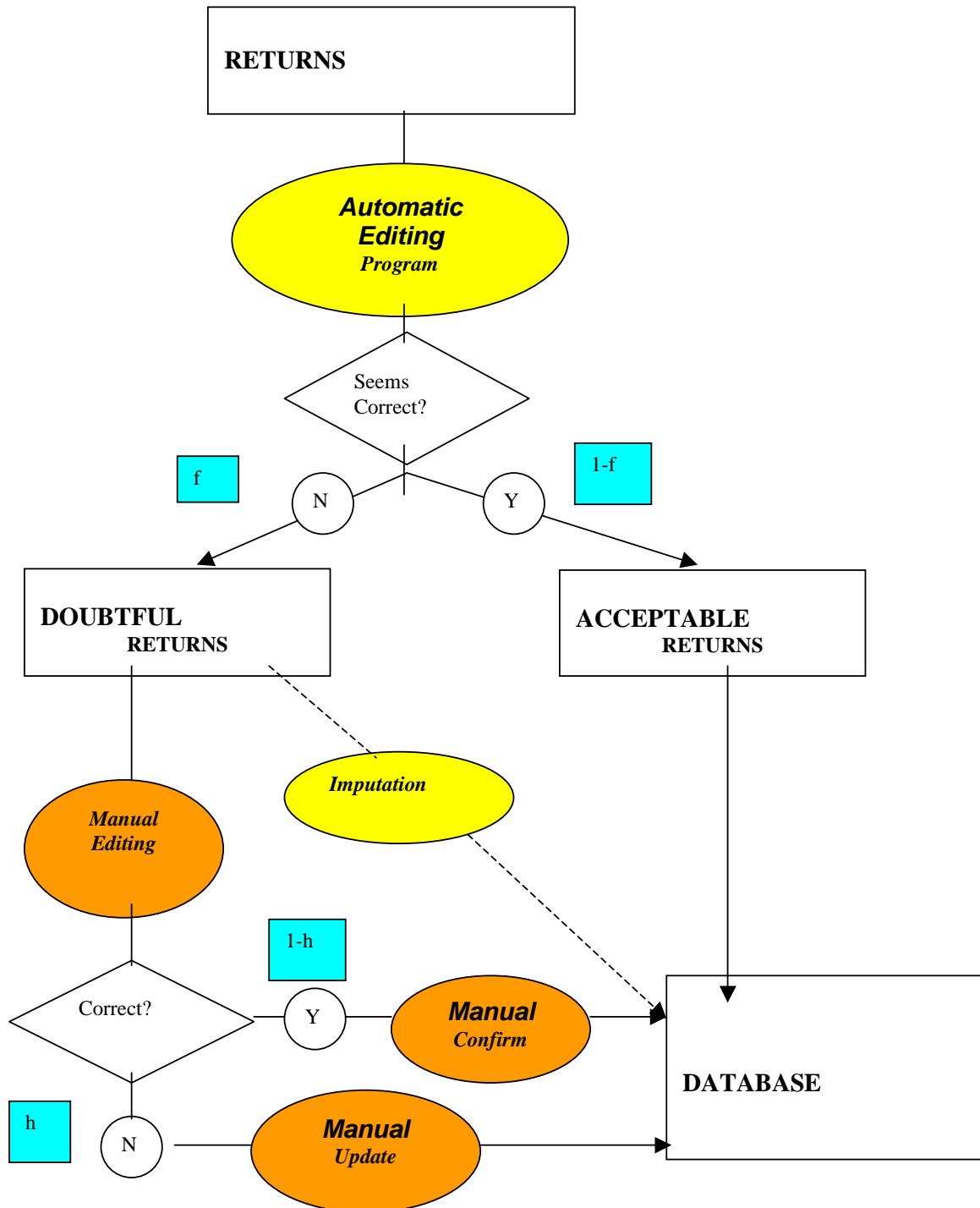
7. Why is it in practice so difficult to optimize data editing in business surveys? The main reason is that the data editing process is far more complex than it seems to be at first glance.

8. First of all, it is a process whose objectives may be unclear: accuracy of the estimates (at various levels of aggregation), accuracy of individual data, low error rate? Data editing can take place at different stages of the process: during data collection, immediately after data collection, or very late. The separation between these various stages is not always simple. Lots of techniques can be used (microediting, macroediting, graphical editing, etc.) depending on the chosen goals, and different ways of optimizing are available. However, it is impossible to optimize everything in the same time: there is a trade-off between many different goals. There must be some kind of metric (see section VII.A for an example of such a metric).

9. Moreover, optimizing would also mean changing the organisation and contents of clerical work (Tate & al 2001). Last but not least, there are many special features of business surveys (Rivière 2001a) that impact the approach of editing: possibility to check with the respondent, complexity of the edits because of interrelated quantitative variables (plausibility edits and not logical edits), handling for particular cases (restructurings, very large firms, etc.), need for coherence with other statistical results.

10. This complexity necessitates a better understanding of the actual data editing process, as it is difficult to optimize a process that one does not know well. That is why this paper starts with a first view of the process: definitions of the main concepts, simplified scheme of the process, and what the components of a data editing strategy should be. This will enable us to have a better idea of what can be optimized. Then in the following sections we will study in more detail each of these 5 components: selection of the doubtful units, prioritization, criterion to stop data editing, organisation of manual editing, use of data editing to improve the next survey process. Given this deeper analysis, we will try to find out what are the various ways of reducing data editing costs without adverse effects on quality. The last section is about how to put theory into practice, particularly in complex business surveys.

11. What is data editing?



**Fig. 1: simple description of the data editing process**

12. In principle the data editing process is simple to describe: an error detection program determines for each return whether it is acceptable or not. If it is acceptable, it goes into the database. If not, it has to be checked manually. In that case, for each variable of the return, if the contributor's value is valid, the survey clerk confirms it (and modifies it if not).

13. The reality is more intricate. In the following, we are going to look at the data editing process in detail. First, as we will see, the concept of validity is not straightforward, and can have several meanings. Second, it is important to be aware that data editing can take place at different stages of the process. Third, improving data editing requires defining a global strategy, whose components will have to be designed efficiently if we really want to improve the process.

#### A. Different notions of validity and major indicators related to them

14. Data editing aims at making sure that data are valid, to a certain extent. The figure points out that distinguishing between “acceptable” and “doubtful” is a decision taken by the computer. There is some test assessing whether a return is acceptable or not, for example using ranges of values that look plausible. Such an assessment is necessarily imperfect: in practice, many doubtful returns turn out to contain no mistakes. On the other hand, the distinction between “correct” and “erroneous” is made by survey operators, and is then the result of a human decision.

15. The concept of validity is therefore non-trivial, as there are at least three levels: **right/wrong, correct/erroneous, acceptable/doubtful**.

16. A **right** value (*the* right value) is the value that corresponds to the reality, the “true” value. This raises philosophical issues that we won’t try to tackle (is there a truth?). Even if we suppose that there is always a right value, it is very difficult in practice to ensure that a value is right: the respondent can lie, sometimes he just doesn’t have the information, or he only has an inaccurate estimation, etc. We will simply say that the *right* value is the value that would be obtained with an *ideal* measurement process (for example: questions asked by a subject-matter expert, best possible respondent, no lies, no restrictions in terms of time or budget, etc.).

17. A value is **correct** if a subject-matter specialist would agree to leave it unchanged in the database. If not, it is said to be erroneous. Correct doesn’t mean “right”. One can leave a record unchanged even if we perfectly know that it is not right: for example if it is impossible to get the true data, or if it is too costly, particularly if we know that the record has a low impact on the statistics that will be produced. But as it would be too expensive to have all questionnaires checked by subject-matter statisticians, a first filtering is done automatically by an editing program. The role of this program is to determine whether the data look coherent and plausible or not, given all the available information.

18. Then a value is said to be **acceptable** if the data editing program accepts it in the database without any requirement to check it manually. If not, it is said to be doubtful. Defining what is acceptable and what is doubtful requires subject-matter expertise about the relationships between variables, about the variability of these relationships, depending on subpopulations (industries, sizes), about the meaning of item non-response for some variables, ... The process of deciding whether a form is doubtful or acceptable is usually combined with identifying where some human input is needed for other reasons such as coding written-in answers. Hence something that is not an error might require attention (manual check), and it is then “doubtful” by definition.

19. Obviously, as the program doesn’t have human expertise, “acceptable” and “correct” are different concepts. By definition, acceptability can be defined formally, correctness (based on human expertise) can not. Therefore, we have the following scheme, similar to (Manzari & Della Rocca 2000):

	Doubtful	Acceptable
Erroneous	a	b
Correct	c	d

where  $a+b+c+d$  is the total number of units.

20. The **failure rate**  $f$  is the proportion of units that do not pass the edits, i.e. the proportion of doubtful units:  $f = \frac{a+c}{a+b+c+d}$

Hence  $f$  is the proportion of returns that will have to be analysed by survey clerks. Therefore, the higher  $f$ , the more expensive the editing. This indicator, that can be easily calculated, does not take into account the actual validity of the returns. It is a proportion of doubtful ones, *according to the editing program*.

21. The **error rate**  $e$  is the proportion of erroneous units:  $e = \frac{a+b}{a+b+c+d}$

Contrary to the failure rate, the error rate is impossible to calculate (except if a subject-matter specialist checks all the questionnaires). It cares only about the correctness of the records, and is totally independent of the data editing process.

22. The **hit rate**  $h$  is the proportion of doubtful units that are erroneous:  $h = \frac{a}{a+c}$

It represents, to some extent, the efficiency of the selection of doubtful returns. If  $h$  is small, it means that many returns were unnecessarily checked. The hit rate, easy to calculate, measures a kind of recovery between “doubtful” and “erroneous”, which means that it takes into account the actual validity according to the statistician and the validity according to the program.

23. After the editing process, we know  $a$ ,  $c$ , and  $b+d$ , but we do not know the division between correct and erroneous among acceptable units ( $b$  and  $d$ ). That is why failure rate and hit rate can be calculated, whereas error rate can not.

24. In many surveys the failure rate is high, and the hit rate is low, which means that too many questionnaires are manually checked. (Tate & al 2001) indicates that in the major ONS business surveys, the failure rate is in the range of 30% to 80%, whereas the hit rate is generally around 20%. When the failure rate is high, as  $a+c$  is high,  $c$  is high and  $b$  is very small, probably negligible. A way of reducing the amount of manual editing is simply to reduce  $a+c$ , which leads to lower  $c$ , ensuring at the same time that  $b$  does not increase significantly. At first glance, optimizing the data editing process could be understood as a sort of trade-off between  $b$  and  $c$ , between checking too many questionnaires and increasing the error. But as we will see, it is only one of the multiple ways of improvement.

## B. Stages of the data editing process

25. In fact data editing is slightly more complicated than shown in the figure. Not only can it be manual or automatic, but it can also take place at various stages of the process. The main steps are:

- early editing
- automatic editing
- current manual editing
- late manual editing

26. Early editing is done during data collection, and is to some extent incorporated to data collection. It comprises many aspects: verifications done by computer programs in a CATI/CAPI/CASI approach, checks associated to keying or scanning programs, or any kind of check or correction done by the surveyor in a face-to-face or a telephone interview (Lepp & Linacre 1993).

27. Automatic editing is performed after the filled-in questionnaires were keyed or scanned. It allows the returns to be broken down into two categories: doubtful ones and acceptable ones.

28. In many NSIs, current manual editing is the day-to-day work of survey operators, in any kind of survey. Returns are manually verified as soon as they are edited automatically, as shown in figure 1. Then doubtful data are verified. A significant proportion of time and resources is spent on these tasks (Tate & al 2001).

29. Late manual editing is probably the least known part of the process. Even if all doubtful data are checked (and corrected if necessary), it turns out in practice that subject-matter statisticians analyse the results from a statistical point of view, and that they do not feel comfortable releasing statistics that do not seem consistent with other published results that should be similar (for example previous statistics on the same subject). It is particularly the case in business surveys, because of the existence of other sources of information. Then if statistical results in a given domain do not look plausible (because of a large change in aggregates, for example), statisticians check the most influential units, or ask survey clerks to do so. In complex business surveys, this late editing requires a lot of work (a couple of months in French annual business surveys, for example) because there are a lot of target variables and target domains. This means that there might be additional validation tests, that can contribute to the low hit rate of editing (Underwood & al 2001). Late editing is not very well formalised, contrary to current manual editing or automatic editing.

### C. Components of a data editing strategy

30. Given this representation of data editing, what do we have to think about if we want to design a “good” data editing system? What do we have to put the stress on?

#### Context

31. Data editing strategy takes place in a framework characterized by the **goals** of the editing process, its **costs**, and its **resources**. The goals of the survey process depend on the **users’ needs**, and the users’ needs themselves can mainly be described in terms of: target variables, target domains (generally industries, regions, size bands, or combination of them), target statistics, the desired level of accuracy, and timeliness.

32. As far as the objectives are concerned, it is important to make a clear distinction between the survey process and the data editing process. The data editing process aims at obtaining records that are sufficiently acceptable, the meaning of “sufficiently” depending on the users’ needs. Formally, these objectives can be defined in terms of coefficients of variation (including measurement error and sampling error), or in terms of error rates, in both cases for each target variable and each target domain. Combining these variation coefficients (or error rates), it is possible to calculate a global score, and to define a minimum score to reach.

33. The total cost of editing, for a given survey, can be considered as the sum, for all the manually checked units, of the elementary costs of checking (and correcting) each of these units. Therefore the set of actually edited units, which is itself included in the set of doubtful units (if we neglect late editing at that stage), will determine the total cost.

34. The resources can be defined in terms of: number of survey operators in full-time equivalent ( $\rightarrow$  maximum number of questionnaires to be edited), available workforce devoted to IT, financial resources, hardware & software. The qualitative side of the resources is also important, as a lot of training is often necessary in order to handle for the responses to complex economic questionnaires, particularly in very large companies.

#### Model

35.  $r$  is the “population”, which is in our case the set of respondents. Let  $N$  be the number of respondents. By definition, respondents returned a filled-in questionnaire, that might contain missing items and errors.

Let  $I_k$  refer to a 0-1 variable indicating whether a return  $k$  is edited (1) or not (0), and let  $J_k$  be a 0-1 variable indicating whether a return  $k$  is erroneous (1) or correct (0).

$Y_k$  will represent the set of values for unit  $k$ .

Let  $P(J_k = 1 / Y_k) = q_k$ , probability for  $k$  of being erroneous, given its set of values.

Let  $e_k$  be a random variable representing the potential error made for a unit  $k$  and a given target variable. Then the error for unit  $k$  will be  $J_k e_k$ .

We will assume that  $e_k, k=1, \dots, N$  are independent,  $J_k, k=1, \dots, N$  are independent, and the independence between all  $e_k$  and  $J_k$ .

It will be assumed that  $V(e_k) = \sigma^2 S_k$ , where  $S_k$  is a size variable for unit  $k$  (for example turnover). The underlying idea is that the larger the size, the larger the error.

We will also consider, only for convenience, that:  $E(e_k) = 0$ .

Such a hypothesis does not prove exact in reality, but it will simplify the expressions, without changing the main ideas: we would obtain similar conclusions with a more general model, and our aim is not to apply the model, to estimate it, but just to draw some conclusions from it.

Therefore, the total measurement error can be expressed as:

$$Err = \sum_{k \in r} J_k e_k (1 - I_k) = \sum_{k \in nc} J_k e_k, \text{ where } nc \text{ is the set of respondents whose questionnaires were not}$$

manually checked. We consider that whenever a unit is checked, all its errors are fixed (which is clearly an approximation).

By hypothesis,  $E(Err|Y) = 0$ .

It is easy to find that:

$$V(Err|Y) = \sigma^2 \sum_{k \in nc} q_k S_k = \sigma^2 \sum_{k \in r} q_k S_k (1 - I_k)$$

On the other hand, checking the questionnaires is costly, and we will assume that the cost of verifying a return is a constant  $c$ . Therefore the total cost of checking is:

$$Cost = c \sum_{k \in r} I_k$$

Optimizing data editing generally means “reducing the costs without adverse effects on quality”, or “improving the quality for a given cost”. Using our simple model, it means minimising the variance of the error (conditionally to the observed values) for a given cost, or minimising the cost for a given variance of the error, which are two dual problems.

$$Min \left[ \sigma^2 \sum_{k \in r} q_k S_k (1 - I_k) \right] \text{ under } c \sum_{k \in r} I_k \leq c^*$$

or

$$Min \left[ c \sum_{k \in r} I_k \right] \text{ under } \sigma^2 \sum_{k \in r} q_k S_k (1 - I_k) \leq v^*$$

where  $c^*$  and  $v^*$  are constants (maximum cost and maximum variance)

How to optimize, according to this model?

It is more complex than it looks (a constrained minimisation problem). For example, one of the main issues is that the  $q_k$  are difficult to estimate.

But in some cases it turns out that  $q_k$  can be considered as very close to 0. This is the case when data are very coherent: *given the set of values  $Y_k$* , the probability of error is very small in that situation.

36. Therefore the first idea is to *eliminate from the editing process the units with very small  $q_k$* . This is exactly the principle of distinction between acceptable returns and doubtful returns. From the model, we can also derive a fundamental idea: **“low conditional probability of error” and “low size” play exactly the same role**. Only the product  $q_k \cdot S_k$  is important. For example, we can “accept” a return whose probability of being erroneous is non negligible, when its size is very small. This is a way of justifying macroediting, and data editing approaches based on impact on aggregates.

37. After restricting ourselves to the doubtful units, a possible idea is to *solve the first constrained minimisation problem*, for a given cost  $c.p$  ( $p$  integer), in variables  $I_k$ . This is a trivial optimization problem, whose result is obvious: the selected units, that will be checked manually, are the  $p$  ones with the highest  $q_k S_k$ . It means that ideally, there will be a prioritization of returns by decreasing  $q_k S_k$ . If we do not have any information on  $q_k$ , it will be considered as a constant, which implies a prioritization by decreasing size. The interesting property here is that the ranking of units will be the same whatever the maximum cost fixed, which is useful as it is very difficult to define a maximum cost.

38. Symmetrically, we can *consider the second constrained minimisation problem*, which lead to stop data editing whenever the obtained accuracy is sufficient. It avoids some unnecessary additional checking, and hence automatically reduces the costs.

39. In the previous approaches, the parameters were considered as fixed, but we can act on them. Therefore another method consists in *optimizing the parameters*: reducing the elementary cost  $c$ , or reducing the error parameters  $q_k$  and  $\sigma^2$ .

### Strategy

40. The ways of optimizing the editing that were derived from our model enable us to suggest a “general data editing strategy”. This strategy will then be characterized by 5 components: automatic editing & imputation, prioritization, stopping, manual editing, process redesign.

- As far as automatic treatments are concerned, the **editing & imputation** program does two things: **selection** of the doubtful returns (by distinguishing them from the acceptable ones), and **imputation** of some variables, this imputation being closely related to data editing. Only a part of the imputation is done at that stage; for example total non-response imputation generally takes place later. The selection of doubtful units obviously has the major role in the process, as minimising the number of doubtful units will directly impact the costs.
- Given a list of returned questionnaires to check, we saw it was useful to rank them, with respect to their size and probability of error, and thus to their potential impact on statistics. Such a **prioritization** enables one to release statistical results at a coarse-grained level very early. Using prioritization, we can also get a quick understanding about where the main problems are, and have the best possible results if there is any kind of restriction in terms of time or budget.
- Whatever the selection and prioritization techniques, it is important to determine when to stop data editing. Implicitly, the **stopping** criterion is often defined as: data editing is stopped when all the doubtful questionnaires have been checked and (if necessary) corrected. This means that a lot of returns are unnecessarily checked. Tuning a stopping criterion is then a way of reducing cost and delays, at the very end of the data editing process. The idea is to stop editing as soon as the estimated impact of the errors of remaining unedited units is “small”.
- A data editing strategy is also characterised by the actual contents of clerical work, which is very costly in business surveys, as it often requires to get data from the respondent. Therefore it seems essential to have a better understanding of **manual editing** (which is often a “black box”): main

tasks of survey operators, organisation of the work, kinds of information provided by data editing programs about the reasons for doubtfulness, ...

- A data editing strategy mustn't be restricted to one survey at one period of time. It has to be broader. An important idea is to take advantage of all kinds of data editing indicators in order to see what goes wrong, which enables one to **redesign the survey process**. The principle here is to prevent errors instead of correcting them.

41. In the following, we will elaborate on these different aspects, in order to get a deeper understanding of what is data editing, and in the same time to have some clues to find how to optimize the whole process. Our main question will then be: how to define a selection criterion and how to connect it with imputation? How to prioritize manual editing? When to stop? What are the main actions performed by survey operators and how to improve their work? How is it possible to redesign the survey process, and what kinds of quality indicators are helpful to find what to improve?

## II. SELECTION OF DOUBTFUL UNITS

### Some notations

42. In the following,  $s$  is the set of units that are sent a form,  $n$  the number of variables in the questionnaire, and  $m$  the number of additional variables that are used for the purpose of data editing.

$Y$  will be the vector of questionnaire variables,  $Y = (y_1, \dots, y_n)$ , and  $X$  the vector of additional variables,  $X = (x_1, \dots, x_m)$ . The additional variables are generally previous values of questionnaire variables, register variables, or variables coming from administrative sources.

For each questionnaire variable  $y_i$  we can define the domain of all possible values  $A_i$ , and we could identically define a domain  $B_j$  for each additional variable  $x_j$ . These domains are generally simple to define: generally  $[0, +\infty[$ , or  $[0, 1]$  for a proportion.

The overall domain for the vector of variables  $Y$  is then  $\prod_{i=1}^n A_i$ .

The overall domain for  $(Y, X)$  is  $Z = \prod_{i=1}^n A_i \times \prod_{j=1}^m B_j$ .

In the following the additional variables  $X$  will be considered to be given (exogenous, to some extent), which means that we can restrict ourselves to the variables of the questionnaire. In other terms, the fact that an external variable might be suspicious is not an issue that will be tackled here.

For a given unit  $k$ , we will denote respectively  $Y_k$  and  $X_k$  the values that the vectors  $Y$  and  $X$  take for the unit  $k$ .

### A. Levels of acceptability

43. How to distinguish between acceptable and doubtful units? To begin with, let us first notice that the concept of acceptability is not straightforward, as it depends on what it refers to. In fact, there are **3 levels** with which we can associate acceptability or doubtfulness: the edit rule, the variable, and the whole return. For a given record, data can fail the edit rule or pass it: for example if the order of magnitude of a ratio is suspicious, we are in presence of an edit rule that fails. We can also assess whether a variable  $x$  is acceptable or not, and this assessment is generally based on the success/failure of the edit rules that take account of  $x$ . Deciding if the return as a whole is acceptable or not is another issue, and it is also often based on the acceptability of each of its variables.

44. Distinguishing these three levels is a way of saying that a return might be considered as acceptable even if some of its variables are doubtful, and that a variable itself might be considered as acceptable even if some edit rules referring to it fail.

45. This way of thinking is very different from the Fellegi-Holt (1976, pp 17-35) approach: “A record which passes **all** the stated edits is said to be a “clean” record, not in need for any correction. Conversely, a record which fails **any** of the edits is in need for some corrections.” With such an approach there is no need for distinguish between three different levels. But it is important to notice that the Fellegi-Holt approach was designed mainly for household surveys, in a context in which the aim is to automate data editing and imputation. Household data are mostly categorical and there is generally a clear distinction between what is consistent and what is not.

46. The situation in business surveys is different, as such a distinction cannot be made: statisticians have to handle “grey areas”. Moreover, it would be extremely risky to automate data editing and imputation: the population of businesses is so heterogeneous that some companies can have a very high impact on the results (Rivière 2001a). Therefore too much automation can lead in some cases to big mistakes ... and eventually to a lot of manual review. Hence the main issue in business surveys is not to optimize an automatic approach, but to find means of reducing the amount of clerical work.

## B. Edit rule level

47. In order to check whether a variable is plausible, the idea is to base a “credibility test” on existing relationships that involve this variable. For example we can compare it with other data of the same unit (register data, survey data, previous data, etc.) so as to assess the consistency of the corresponding set of values. Edit rules (or “edits”) are then plausibility criteria involving several variables. In business surveys, the criterion is generally a ratio belonging to an interval, as in household surveys, it can often be a nested if-then-else statement.

48. There are two kinds of edit rules: query edit rules and fatal edit rules. If a unit fails a fatal edit rule, it means that the error is certain. On the contrary, query edit rules identify data that are not plausible: even if they are well designed, the probability of error is often far from 100%.

### Definitions

49. Our aim here is to define an edit rule mathematically. First of all, an edit rule enables to split the domain of possible values into two parts: the ones that pass the edit and the ones that does not.

An *edit rule (or edit)*  $e_p$  is a function of  $Y$  and  $X$  that defines a subset  $Z_p$  of the overall  $(Y,X)$ -domain,  $Z$ :

$$Z_p \subset Z = \prod_{i=1}^n A_i \times \prod_{j=1}^m B_j .$$

Then a unit  $k$  will be said to pass the edit  $e_p$  if and only if  $(Y_k, X_k) \in Z_p$

For a given unit  $k$ ,  $X=X_k$  is known, which enables to define, for every edit rule  $e_p$ , the domain associated

to it as subset of the overall  $Y$ -domain  $\prod_{i=1}^n A_i$ , depending on  $X_k$ :

$$E_{pk} = \left\{ Y \in \prod_{i=1}^n A_i \text{ such that } (Y, X_k) \in Z_p \right\}$$

Now we want to explain how to build such an edit rule. In practice, in many cases, it is defined by a ratio of 2 variables, and an interval. The unit passes the edit if its corresponding ratio falls in the interval, and it fails the edit if not. Because of the behaviours of companies vary a lot, depending on industry or size, there can be several intervals, each one associated with a subpopulation.

Therefore, we can define an *edit rule*  $e_p = \{f_p, C_p, d_p\}$  as:

- an *edit function*  $f_p(Y, X)$  that takes its values in  $\mathfrak{R}$ , and that depends on one questionnaire variable at least<sup>2</sup>.
- a *finite set of categories*  $C_p = \{c_{p1}, \dots, c_{pn_{pi}}\}$ , such as each unit in  $s$  belongs to a category and only one category (examples: industry, size band)
- a *validity domain function*  $d_p$  defined on  $C_p$ : each category  $c_{pj}$  has its validity domain, which is a set of real numbers, generally an interval.

As the categories have to be computable for every unit, using  $Y$  and  $X$ , we shall denote  $q_p(Y, X)$  the categorising function, which takes its values in  $C_p$ . These categories are a partition: every unit belongs to one category, and only one.

A record will be said to *pass the edit*  $e_p$  if and only if its variables  $(Y, X)$  are such that

$f_p(Y, X) \in d_p(q_p(Y, X))$ . If not, it will be said to *fail the edit*. This means that the *acceptability domain*  $Z_p$  associated to the edit  $e_p$  can be defined as:

$$Z_p = \{(Y, X) \in Z \text{ such that } f_p(Y, X) \in d_p(q_p(Y, X))\}$$

In a data editing system, the *edit set*  $\Lambda = \{e_1, \dots, e_p\}$  is the finite set of all edit rules.

### **Example**

50. If we want to check the validity of the variable Turnover, a possible edit function would be Turnover / Turnover(t-1). The variable Turnover belongs to the questionnaire and then belongs to  $X$ , as the variable Turnover(t-1) is an additional variable.

This ratio can vary a lot from an economic sector to another. Thus the set of categories  $C_p$  is often a set of sectors, or a set of sectors by range of size. It is important to ensure that the categories have a certain level of homogeneity.

51. Variables used to partition the population, like SIC, number of employees or region, are assumed to be known (if not, it would be impossible to define a partition). Therefore they belong to  $Y$  or to  $X$ , preferably to the vector  $X$  of additional variables, as these variables have to be known before sending the form. Then from that information, it will be possible to find which category a given unit  $j$  belongs to: that is the meaning of the function  $q_p$ .

52. For each of those sectors, validity domains will be intervals, like  $[0.7 ; 1.3]$  in the case of Turnover/Turnover(t-1). For a given edit, each category is then associated to a validity interval. (Thompson & Sigman 1999) propose statistical methods for automatically setting these tolerance limits.

### **Some principles to design edit rules**

53. Few variables are used in a given edit rule: generally between two and four in the edit function  $f_p$ , and two or three in the definition of the categories,  $q_p$ .

---

<sup>2</sup> If not, the edit would not depend on questionnaire variables, therefore it would be irrelevant for the purpose of editing.

54. For purpose of simplicity, the categories are often the same for all edits, which means that  $q_k$  is independent from  $k$ . If not it can be a little bit difficult to manage categories in the editing software. In business surveys, a category is generally an industry, or a size band, or the combination of both.

55. For the same reason, the edit functions will often be the same across the categories, but the validity domains may vary from a category to another. If an edit rule is really specific to a particular category  $j$ , it can be represented with the approach proposed: for all categories but  $j$ , the acceptance regions will be  $]-\infty, +\infty[$ .

56. One of the most difficult things is to define the edit function. In the case of quantitative variables, they are generally ratios of two “similar” variables. To find out what “similar” means, it is helpful to define a variable by a concept, a source and a period, the idea is to compare it with another one that differs from it on one point.

Type 1: another concept, same source, same period. Then finding the “other concept” is another issue, which often requires some background in accounting.

Type 2: same concept, another source (administrative for example), same period. The idea is to compare with a value that should be the same.

Type 3: same concept, same source (the survey), another period. The ratio is then often a growth rate.

### **Edit rules in business surveys**

57. In business surveys, data are generally quantitative. Then edit rules are generally ratios, and validity domains are ranges of values (Granquist 1995, Thompson & Sigman 1999), positive in most of the cases. A majority of the edit rules are query edit rules. For a given variable of the questionnaire, there might be several edit rules, as there are many relationships between economic variables. The fact that the majority of business surveys are repeated surveys also helps to define edits.

58. In practice we can distinguish between three kinds of edit rules in business surveys (Rivière 1997): ratio edits, time edits and balance edits. A **ratio edit** is an edit such as the editing function  $f_p(X, Y)$  is a ratio of two variables corresponding to the same period. It can be two variables of the questionnaire, for example value added and investment, or turnover and number of employees; this is then a ratio edit of type 1.

Ratio edits of type 2 can consist in dividing a variable by the corresponding variable in the register. We can also compare a variable of the questionnaire with a variable coming from an administrative source. However, it is important to stress that administrative data are not that easy to use. They contain many inconsistencies, which are not necessary “errors”, but “anomalies”, and the semantic differences between administrative concepts and survey concepts can play a key role. See (Boydens & al 1997, Boydens 1999) for a detailed discussion of the issues raised by administrative data.

59. A **time edit** is an edit in which a given variable (or ratio) is compared to the same in the previous period (= ratio edit of type 3). It is possible to apply such an edit only if previous values are available, which is not always the case because of sample rotation.

Example:  $\text{turnover}(t)/\text{turnover}(t-1)$

60. A **balance edit** is an edit in which the formula corresponds to an accounting equation like  $T = \sum_i T_i$  (for example  $T$  is the total purchase, and the  $T_i$  are the different kinds of purchases).

The main difference between this kind of edit and the two others is that such an equation is exact, as other edits corresponded to “plausibility”. In practice balance edits admit a slight error, and are written as :

$$\left| T - \sum_i T_i \right| \leq a, \text{ if } T = \sum_i T_i \text{ is the accounting equation.}$$

Therefore balance edits are far more reliable than the other edits. Experience also shows that time edits prove better than ratio edits.

### Significance edits

61. The previous edits aimed at finding inconsistencies (within the questionnaire, with previous data, or with other sources). That is a way of selecting the doubtful units. However, it is important to recall that the aim of automatic editing is not to have true values, but acceptable values, which means “values that can be accepted without any manual checking”. Therefore it can prove useful to verify data that have a huge impact on aggregates, whatever their internal consistency. This approach is radically different. It means that there are some *significance edits* (Latouche-Berthelot 1992, Lawrence-McDavitt 1994) that do not try to identify any kind of inconsistency, but just assess the impact of a given variable (or ratio) on the corresponding aggregate (or ratio of aggregates). The edit function is then often a ratio, in which the denominator is an aggregate. It can be more complicated if the aim is to measure the impact on a ratio, or an index. Mathematically, it is very similar to the previous kinds of ratios, except that we do not divide the variable by a “similar one” (neither type 1 nor type 2 or 3). But the underlying philosophy is completely different.

62. These techniques can prove extremely efficient on business surveys, particularly when the variables are not too numerous. In the case of ONS’ Monthly Inquiry into Distribution and Services Sector (one or two key variables), (Underwood 2001) shows that efficiency savings generated by this approach are very high, as the reduction in the numbers of questionnaire that would have been followed up during editing is approximately 34%, without adversely affecting the quality of the outputs.

### C. Acceptability of a variable

63. In practice, the fact that a unit fails an edit or not is not necessarily an important question in itself. It is different for variables: if a variable is doubtful, a decision has to be taken on whether to impute it or not. The definition of the acceptability of a variable can be based on formally defined edit rules, but it is not necessary (graphical editing for example). If it is the case, it is clear that the acceptability criterion will be a function of all the edit rules that involve this variable, and that there will be at least one edit rule for each variable we want to edit. In an edit-rule-based system, the main questions are then: how to combine the edit rules in order to get the criterion for the variable? If the variable is doubtful, what to do with the value?

### Definitions and properties

64. First of all, the *acceptability domain associated with a variable*  $y_q$  is a subset  $Z_{(q)}$  of the initial acceptability domain:  $Z_{(q)} \subset Z$ .

Then, by definition, for a given unit  $k$ , the variable  $y_q$  will be said to be acceptable if and only if

$$(Y_k, X_k) \in Z_{(q)}.$$

If the acceptability of a variable is based on a combination of edit rules, what we want to do is to establish a relationship between the acceptability domain of the variable and those associated with the edit rules.

Hence it is important to find in which edit rules the variable  $y_q$  plays a role.

An edit  $e_p = \{f_p, C_p, d_p\}$  is said to *depend on* a variable  $y_q \in Y$  iff the edit function  $f_p$  depends on  $y_q$ .

For every  $y_q \in Y$ , the *edit set* of  $y_q$  is the set  $e_{(q)} \subset \Lambda$  of all edits depending on  $y_q$ .

An *unedited variable* is a variable  $y$  such that  $e_{(q)} = \emptyset$ . By definition, such a variable is always considered as acceptable whatever its value.

65. We can assume the two following properties:

If the variable passes all edit rules, then it is acceptable.

If the variable is acceptable, then it passes at least one edit rule.

The situation where the variable  $y$  passes all its edit rules corresponds to the following acceptability

$$\text{domain: } Z_{(q)}^{\min} = \bigcap_{\lambda \in e_{(q)}} E(\lambda)$$

This is the most natural way of defining the acceptability of a variable, the simplest one, which is generally used in practice.

On the other hand, passing at least one edit rule means being in the following domain:  $Z_{(q)}^{\max} = \bigcup_{\lambda \in e_{(q)}} E(\lambda)$

Thus we have:  $Z_{(q)}^{\min} \subset Z_{(q)} \subset Z_{(q)}^{\max}$ , which gives a lower bound and an upper bound to the acceptability domain of the variable  $y_q$ .

It is important to keep in mind that the acceptability domain  $Z_{(q)}$  associated with a variable  $y_q$  is not a

subset of its domain  $A_q$ , but a subset of  $Z = \prod_{i=1}^n A_i \times \prod_{j=1}^m B_j$ .

But given all the other  $y_i$  and given  $X$ , it is still possible to determine a *conditional acceptability domain* of  $y_q$ :

$$A_q(y_1, \dots, y_{q-1}, y_{q+1}, \dots, y_n, X) = \left\{ a \in A_p \text{ such that } (y_1, \dots, y_{q-1}, a, y_{q+1}, \dots, y_n, X) \in Z_{(q)} \right\}$$

### Building acceptance criteria for variables

66. In many editing systems, a variable is considered as acceptable only if it passes all the edits that involve this variable. This is a defensive approach, which corresponds to the lower bound of the acceptability domain. It is not a very efficient approach in terms of reduction of the costs: *the more numerous the edit rules, the less chances you have to accept a value*. It is then clear that it increases the chances of doubtfulness. Moreover, it happens sometimes that only one variable in a return is erroneous; in that case all edits with this variable will fail, which will automatically lead to finding doubtful all the variables that appear in these edits.

67. Another approach consists in combining the edit rules in a less defensive way. For example if a variable is involved in 4 edit rules, and if it fails only one of them, it might be reasonable in some cases to consider that this variable is acceptable.

68. But the acceptability is not only a function of numbers of failures of the edits: the fact that each edit rule gives only two possibilities (failure or success) is a limitation. It is more efficient to define *scores*, which will generally be kinds of distances between ratios and intervals.

Reusing notations of 2.b, a score associated with edit  $e_p$  will then be defined as:

$$S_p(Y, X) = D(f_p(Y, X), d_p(q_p(Y, X)))$$

where  $D$  is a distance (between a number and a set)

69. Then combining the edit scores of a variable will provide a global score, which will serve as a basis to define the acceptability criterion we need: for example if the global score is lower than a threshold, the variable is considered as acceptable, if not, it is said to be doubtful.

70. This kind of approach has been applied to the French annual business surveys (Rivière 1997). The idea is to have two plausibility intervals per edit, a small one and a large one. Then for each edit there are three marks: 0 if it the value falls out of the large interval,  $n_1$  if it falls in the small one, and  $n_2$  otherwise. The marks are different for each edit, but we always have  $n_2 < n_1$ . Then a combination of the marks give a score for the variable. The way of combining the marks differs from one variable to another, as the number of edits and their relevance vary a lot.

71. One of the main difficulties with this kind of approach is that the corresponding domain is far more difficult to visualise and to maintain. Moreover, it is not convex, therefore it is impossible to apply Chernikova-like algorithms (cf. Chernikova, 1965 for the theory; De Waal, 1997, for the application to editing). But if the edit rules and the rules of combination of scores are well documented, this technique can reduce substantially the amount of editing.

### Missing values and written-ins to be coded

72. Our suggested representation of the problem was incomplete, as we should take into account the missing values. Technically: each variable can take a particular value named *missing value*, denoted  $M$ , which means that the variable was not filled-in by the respondent.

73. One could think that a missing value should be imputed, but it is less simple than that. It can happen for example that there are different questionnaires (depending on the industry), which implies that some variables are not present in all questionnaires: the missingness is somehow normal (represented by a “null value” in the database). This can also happen even if the variable is present in the questionnaire.

It implies that for every variable  $y_q$ , it is possible to define a set of units  $s_q$  in which  $y_q$  is *relevant*.

For a given unit  $k$ , if  $k$  is in  $s_q$  and if  $y_q$  is missing, then  $y_q$  is obviously “doubtful”, which is equivalent to say that something has to be done for this variable.

74. The existence of a variable with a missing value also has an impact on the variables involved in edits with the missing one, as they will have a higher probability of being seen as doubtful (and a probability 1 if the “defensive” approach is used). That is one of the reasons why it is important to make the editing process flexible.

75. Finally, the non-coded written-ins (for example the description of an economic activity) are in a way a particular case of missing values: the relevant value (e.g. the code) is missing. It is clear that the missing code behaves like a doubtful value. A way of tackling this problem is to apply automatic coding.

### D. Acceptability of a return

76. The main concept of acceptability is this. The acceptability of the variable is important because it can lead to imputation, but the acceptability of the return is absolutely essential: it is on the basis of this acceptability that we decide whether the return will require manual work or not. The main problem is that the cost of manual data editing does not seem to be proportional (even approximately) to the number of *values* checked manually. It appears that this cost is directly linked to the number of *questionnaires* checked. The reason is that whenever a questionnaire is manually verified, it cannot take a few seconds,

as the survey clerk needs to have an overview of all contributor's data in order to understand what is going on, and then to determine where the errors could be.

77. Therefore, in terms of reduction of the costs, it is crucial to have in mind that the main issue is not to know whether a given value passes an edit or not, as the only important thing is to determine whether the entire questionnaire will be considered as plausible or not.

### Definitions

78. The acceptability of a return is defined by a *global acceptability domain*  $Z^*$ , which is a subset of the initial acceptability domain:  $Z^* \subset Z = \prod_{i=1}^n A_i \times \prod_{j=1}^m B_j$ .

A return  $k$  is then considered as acceptable iff  $(Y_k, X_k) \in Z^*$ .

79. This definition is less trivial than it seems. It is important to notice that the external variables can play a role in the acceptability of a return. The reason is that the non-acceptable units are units that we don't want to use as they are. It's not only because of their answers to the questionnaire. For example, in many enterprise surveys, a large company involved in a restructuring will always be verified manually, whatever the contents of its questionnaire, because the probability of error (and of large error) is high, whatever the actual data. It can be the same for other particular cases, well known by survey operators: for example, when companies that fill in a "comment box" (Underwood 2001), survey operators check the comments, which means that they perform a kind of manual editing.

80. When edit rules are defined, the acceptability of the return can be based (at least partially) on the other levels of acceptability (edit level and variable level).

A defensive approach consists in defining the acceptability domain as the intersection of acceptability

domains of all variables of the questionnaire:  $Z_1^* = \bigcap_{q=1}^n Z_{(q)}$ .

Even more defensive, we could take the intersection of all acceptability domains of all edit rules:

$$Z_0^* = \bigcap_{p=1}^P Z_p$$

81. Is it possible to do better? In fact, the main problem is that if we accept a return in which part of the variables are doubtful, it means that we "accept" some doubtful values. This raises the question: What to do when there are doubtful variables? There are 3 possibilities:

- let all values unchanged,
- check the questionnaire manually,
- impute the doubtful values (or at least part of it).

82. The first solution means that the return is considered as acceptable. The second one means that the return is seen as doubtful. The third solution is between the two: the raw return is doubtful, but the "transformed" return (after imputation) is acceptable. This raises the issue of the links between editing and imputation.

### Links between editing and imputation

83. The fact that data editing can lead to imputation of doubtful values which might be "accepted" (without any manual review) implies that there is a notion of "first-order acceptability" and "second-order acceptability".

The first-order acceptability (or “acceptability”) of a return  $k$  means that its set of raw values is accepted without any change:  $(Y_k, X_k) \in Z^*$ . It is the usual definition of the acceptability of a return.

84. The second-order acceptability is less restrictive. A return is said to be 2<sup>nd</sup>-order acceptable if a “transformed” set of values, in which part of the doubtful values are imputed (acceptable values being left unchanged), can be accepted without any manual verification. The important point is this absence of manual verification, as the aim of the process of selection of doubtful units is to minimise the amount of manual checkings.

85. The transformed set of values is not necessarily first-order acceptable. In that case it means that the statistician in charge of the methodology admits that there is an early imputation of “some” variables, without checking the properties of imputed values in terms of internal consistency. Implicitly it also means that the acceptability domain for the transformed set of values is larger than  $Z^*$ . To what extent can imputation be used for doubtful values? It is a difficult question. The problem is to avoid that imputed values have a high impact on the statistics. For example, in the methodology used in the French annual business surveys (Rivière 1997), an imputed value is accepted as long as its distance to the initial value is below a given threshold, but the acceptability of the new values is never checked.

86. An imputation methodology independent from data editing has a main advantage: it generally enables one to control the statistical properties of the aggregate. On the other hand, nothing ensures that it is consistent with the rest of the return, and it can even happen that the imputed value is worse (according to data editing scores) than the raw value. In such a case it might be better to use a mixed technique: if the imputed value improves the score, take the imputed value. If not, keep the initial one.

87. In the Fellegi-Holt (1976) methodology, all imputations are made to ensure that the transformed set of values passes all the edits. The method also minimises the number of fields to impute. A way of seeing this methodology is to say that:

- the definition of the (first-order) acceptability of the return is the most “defensive” one: intersection of acceptability domains of all edit rules (called  $Z_0^*$  previously)
- the returns are always considered as second-order acceptable,
- they are transformed by imputation into first-order acceptable returns

88. (Little and Smith 1987) also propose a method to edit and impute automatically. But, as said before, an automatic editing and imputation method, whatever the method, is not very well adapted to business surveys, for many reasons. One major reason is the fact that edits are generally only query edits (plausibility edits), difficult to tune, such that a return can fail many edits and be perfectly correct in the same time. Therefore making sure that the final vector of values passes all the edits, and hence falls into this domain, is probably not a very relevant objective. Many correct vectors of values would then be forced into this domain, biasing the aggregates. In business surveys, manual review is an essential part of the process. Consequently, even if an automatic editing and imputation method is used, it will not solve the main problem of editing, as it will have to be combined with an interactive checking. For example, (Pierzchala 1995, p.435) stresses that “the SPEER system has combined the Fellegi and Holt methodology with interactive processing whereby the specialist can refer and treat any referred forms”.

### **Techniques with no parameter tuning: graphical editing and outlier detection**

89. Defining edit rules, acceptability domains at various levels, and therefore tuning the parameters, can require a lot of time, and it is not always necessary. We must keep in mind that the main level of acceptability is the level of the whole questionnaire: if we are able to determine this acceptability in a relevant way without writing any edit rule, it is not a problem at all.

90. Graphical editing is typically the kind of technique that gets rid of this tedious work: by visualising the set of values, the statistician can determine which are the most suspicious units. When we are doing microediting, looking at inconsistencies, graphical editing can help to find outliers. Hence no formal edit rule needs to be written, and the acceptability domain at the unit’s level can be seen

graphically. Graphical approaches can also be used to perform macroediting in a visual manner. These techniques have been more and more used during the last decade, as it proved very efficient (De Waal & al. 2000, Weir & al. 1997, Esposito & al. 1997). It has many advantages, for example the fact that the user can in the same time check and correct errors, tune the editing process, and analyse the results (which implies that late editing is no longer necessary in that case). Moreover, (Engström et al. 1995) points out that “graphical applications are not costly to develop”, as long as “the right programming tools and some programming skills” are available.

91. On the other hand, graphical techniques based on exploratory data analysis necessitate that survey operators have a higher level, and noticeably a better background in statistics and information technology, as they have to make technically informant judgments. It is clear that such a requirement increases the costs.

92. When applying graphical editing to business surveys, it is important to take account that characteristics of companies vary a lot from one industry to another, or from one size band to another. A graphical approach is perfectly adequate to handle that problem, with a top-down approach which “provides the user the ability to drill down through the aggregates through the respondent level” (Weir & al. 1997). As it starts from the aggregates and gives information on the individual impacts on statistics, such a graphical approach also corresponds to macroediting.

93. Automatic outlier detection is another approach to find non-acceptable units without defining and tuning edit rules. Indeed, outliers are particular cases of very doubtful units. To some extent, it means that we have one complex edit rule that enables to find abnormal units. (Little & Smith 1987) propose a general methodology that combines detection of outlying cases, detection of outlying values within outlying cases, imputation of likely values for missing and/or outlying and edited values. But such an approach might be difficult to apply when the variables are too numerous and when the population (very heterogeneous) has to be broken down in several subpopulations. Moreover, in business surveys, an outlier is not necessarily an erroneous value, and that is because of the great heterogeneity of the population of businesses.

### **What to do if there are many variables in the questionnaire?**

94. The more the variables, the higher the probability of having at least one error. This probability is trivial to estimate, based on the probability of error of each variable, if all the errors are independent. But it is not the case in practice. (Naus & al 1972) and (Hedlin 2001) propose solutions to estimate this probability, without assuming the independence.

95. Therefore, if the acceptability of a return is based on a defensive approach (intersection of the domains of all edit rules, as in the Fellegi-Holt technique), the proportion of acceptable ones will tend to be small, and therefore the automatic editing program will prove inefficient.

96. A possible solution is to focus on a subset of variables, and implicitly to admit that the verifications on the other ones will be less strict. For example, if one of these variables fails some edits, it can be imputed without any additional check.

97. In order to avoid too much manual verification in the case of large surveys, it is important to bear in mind that acceptability is not a pure concept. It totally depends on the statistics we want to produce, in terms of variables and domains. This means that from one period to another, criteria for selection of doubtful units should change according to the objectives. For example one can imagine that the survey manager defines a list of target variables before each survey, and that a return in which none of the doubtful variables is a target variable is automatically imputed or accepted without further verification. Target domains are also important if the selection criterion is based on impact on statistics: the larger the domain, the smaller the impact, the smaller the chances of being doubtful.

98. In practical situations it seldom happens in complex surveys that data editing criteria can be redefined according to the objectives. It is because tools designed for these complex surveys are inevitably complex, and therefore it is rare that these tools are flexible at the same time.

### III. PRIORITIZATION CRITERIA

99. In the previous section, a return (or a variable) was considered either acceptable or doubtful: 0 or 1. But the frontier between acceptability and doubtfulness is not that trivial. There is a continuum between the two: a doubtful unit can be “more doubtful” than another one. Then it is not  $\{0, 1\}$ , but  $[0, 1]$ . We can use this degree of doubtfulness to prioritize the checkings.

#### A. General ideas

100. There is a great heterogeneity of the errors in business surveys: some are more important to fix than the others. But, as mentioned by (Hedlin 2001), the concept of “important error” depends on the viewpoint: error-based or estimate-based. In the first “error-based” viewpoint, a highly erroneous return is one in which the expected number of errors is high (lots of doubtful variables, with high probability of error). We want to have the cleanest possible dataset. In the second “estimate-based” viewpoint, a highly erroneous return is such that it has a high impact on target statistics.

101. As it is impossible to check everything simultaneously, the editing task will be done in a given order. This is a process that we have to optimize. Prioritizing the actions is a way of optimizing. We saw in the section on data editing strategy, that there was a theoretical background to such a prioritization: minimising the cost of editing for a given “quality” automatically results in a ranking. In the error-based approach units are ranked according to probabilities of error  $q_k$ , as in the estimate-based approach, the sorting is based on the sizes  $S_k$ . We also found out that in theory, units should be ranked according to  $q_k \cdot S_k$ .

#### B. The concept of score

102. A score is a measure of the degree of doubtfulness. It applies at different levels, as for selection: score of an edit rule, score of a variable, score of a return. If we apply the first “error-based” viewpoint, the idea is to use a standardised distance between a value and an interval. For the return level: it is just a combination of scores. Then there is a positive link between the score and some kind of expected (and maybe weighted) number of errors. (Latouche & Berthelot 1986) give a typical example of such a score.

103. If we adopt the second viewpoint, the score at the variable’s level has to be defined as a kind of standardised impact of the unit on a statistical target associated with this variable (aggregate, ratio of aggregates, change in aggregates, coefficient of variation, relative mean squared error, etc.). Then the global score (at the return’s level) can be a combination of variable’s scores, as in previous case.

104. The methods used in practice combine both approaches. In some cases they are based on a previous error-based editing system, to which an estimate-based strategy is added. See (Latouche and Berthelot, 1992, Lawrence and McDavitt, 1994, Lawrence and McKenzie, 2000, Hedlin, 2001).

105. A typical score is the absolute difference between the raw value and an “expected” value, multiplied by the sampling weight (and divided by the aggregate to standardise it). It is clearly a mix of error-based and estimate-based. The fact that the value is compared with a supposed-to-be acceptable value can make us think about a microediting approach (error-based). On the other hand, such an edit rule does not care about the consistency between the variable and the others, and it takes into account the size of the unit, and hence its impact on aggregates (then estimate-based).

106. In a pure estimate-based approach, there would be no “expected” value, which means that the score would simply be the impact on the aggregate. But it would be very inefficient: for example, such a score would not enable us to find returns of large businesses for which some values are zero (because of lack of time to fill-in the form), that obviously have a high influence on the errors.

## C. Issues

107. As we have to take into account the impact on estimates (which means that the definition of scores totally depends on what these estimates are), the viewpoint adopted in business surveys is generally mixed. Therefore, whatever the technique used, the main question to ask before designing (or tuning) a data editing system is: What are the objectives? This question can be rewritten more clearly as: What are the target variables? What are the target domains? What are the target statistics (aggregates, ratios, etc.)? To a certain extent, this is a question for the users. Given the objectives, the main question that the methodologist has to answer is: How to weight the different objectives in an overall global score?

108. This is even more complicated than that, for if we want to estimate an impact on an aggregate, we have to define how to estimate this aggregate. The estimation will vary along the way, as more and more units are checked and modified. We also have to define what is the individual value that will be used to calculate the impact, particularly if this value is doubtful: imputed value? Raw value? The answer of the literature seems to be “the simpler, the better”. Typically, using the previous value as an “expected” value, but also, if necessary, as an imputed value (Lawrence-McKenzie 2000, Hedlin 2001), is a robust approach that can be justified. It is clear that for the purpose of the prioritization of editing, it is no use designing sophisticated models, as the aim is not to estimate anything, but to point out and to rank the units which might have the largest influence on the errors.

## IV. STOPPING CRITERIA

### A. General idea

109. Even if we have an excellent criterion to distinguish between true and wrong questionnaires, is it absolutely necessary to check all the doubtful ones? No! The definition of the data editing decision rule is not sufficient to describe the process in a theoretical way. Using a decision rule to distinguish between acceptable and doubtful records does not mean that all the doubtful will be manually checked. It might be inefficient, and sometimes impossible (for example when statistics are needed quickly).

110. Then, for a given decision rule, one can have an additional decision criterion that tells whether manual editing can be stopped or not. This decision criterion is then a new filter, it is a new selection process in which we take account of the current situation (i.e. the current estimation of target statistics).

111. From a theoretical point of view, it is a fourth level of acceptability: 1) Edit rule, 2) Variable, 3) Return, ... 4) Set of returns. Finding a stopping criterion is equivalent to determine whether a given set of returns is “acceptable” (it can be used directly to compute the statistics) or not.

Theoretically, data editing could be described as:

Is the set of returns acceptable? If yes, stop editing. If no, edit the least acceptable questionnaire.

112. Generally, we stop because:

- 100% done, or 100% done among feasible;
- deadline;
- budget limits (it is quite the same);
- subject-matter statisticians are satisfied with the results (then no formal criterion);
- 100% done for the highest ranges of scores.

### B. What is the use of a stopping criterion?

113. By definition, a stopping criterion has to be used ... to decide when to stop. But such a criterion has other advantages. Even if the criterion is not very relevant, it enables us to have an approximate evaluation of the necessary amount of manual editing (and hence the cost of the survey), even before the survey is launched. If we are able to provide such a measure, it is then possible to define a relationship between the estimated quantity of manual editing and the number of target domains. Then, ideally,

statisticians could say to users: if you want to get statistics by industry 2 digits, it will cost  $x$ , if you want the same quality of results by industry 3 digits, it costs you  $y$ . How much are you prepared to pay?

114. Another advantage is that it enables us to have a good trade-off between current manual editing and late editing: whatever the editing technique applied, late editing will always be used. Therefore it is better not to waste too much time on current manual editing.

### C. Techniques

115. The possible stopping criteria are:

- Measures of impact, but at a macro level (and then thresholds on this measures)
- Remaining error rate: if the estimated proportion of remaining errors is below a given threshold (the target error rate). It is too simplistic, and could hardly be applied on business surveys because of the great heterogeneity of error sizes. But it is useful to get an approximate evaluation of the cost of editing (at least a lower bound). (Rivière 2001b) shows that the minimum proportion of returns to be

$$\text{edited is } \left[ 1 + \frac{rN}{Hg_{\alpha}^2 \left( 1 - 2\frac{r}{f} \right)} \right]^{-1},$$

where  $r$  is the target error rate,  $N$  the number of returns,  $H$  the number of dissemination domains (being a partition),  $f$  the failure rate of the automatic editing process, and  $g_{\alpha}$  the  $1-\alpha$  quantile of the standardised gaussian distribution.

- Current coefficient of variation (or MSE) for each target variable for each target domain, and combination of all: it can provide a score for the purpose of prioritization, and given a threshold it also provides a stopping criterion.

116. Even if prioritization criteria can provide automatically the idea of stopping criterion, both ideas are very different. On one hand, in the case of prioritization, we need a score for each return, and this score is used to prioritize, as units will be checked by decreasing score; but in that case, nothing is said about when to stop. On the other hand, such a score can be used to define a stopping criterion: a threshold is fixed, and as soon as the scores of all returns are below the threshold, the process can be stopped ; but this score is not necessarily used to prioritize.

117. Moreover, a stopping criterion can be defined without using any score at the return level. For example one can decide to stop the process as soon as some variation coefficient is below a target value: it is a global criterion, that does not require the calculation of scores for returns.

118. With prioritization, one defines in which order the elementary tasks will be done: it is about the editing process itself. With a stopping criterion, we do not pay any attention to the process: we have general indicators that tell us whether an adequate accuracy is reached or not, irrespective of what happened before.

## V. MANUAL EDITING

119. As said before, the total cost is somehow proportionate to the cost of manual editing for one return. But generally statisticians do not know very well what manual editing is. They often think that it is simple to describe. Anyone who has interviewed survey operators knows that it is extremely complex, and that it influences the total cost more than we could imagine at first glance. Late editing is one of the examples. Therefore it seems very efficient to have a deeper understanding of what manual editing is.

### A. What does manual editing consist of?

120. Here is a list of tasks of survey operators for the purpose of data editing:

- Viewing the results of a unit
- Checking the validity of doubtful values
- Consulting other sources of information
- Modifying values
- Confirming values
- Coding wordings (wordings are a particular case of missing value in which the survey clerk is able to find the right value)
- Contacting the contributors (and finding the person able to answer the questionnaire, negotiating with them, explaining the survey, etc.)
- Following-up non-respondents
- Feeding a documentary system

121. Non-respondent follow-up is particularly important: if we want to optimize, we have to make a choice between checking questionnaires and contacting non-respondents. This means that any criterion to reduce manual editing has to take account of these two dimensions.

122. There are different levels of complexity in the main task of checking + modifying or confirming:

- simple errors like scanning
- semantic errors that do not require to contact contributors
- errors that require to contact contributors

123. The cost of an elementary verification depends on this complexity: it is clear that correcting scanning errors is far less costly than recontacting a contributor.

## **B. Main difficulties**

124. The actual organisation of survey operators does not ease the optimization of the process. In practice, there are specialized teams, and the repartition is made generally by main industry. Therefore, if it turns out that too much time is spend on activity A1, and not enough on activity A2, optimizing manual editing means that some people of the “A1” team will have to move to the “A2” team, and will also have to get trained on the peculiarities of activity (or group of activities) A2. It is very difficult (and sometimes impossible) to make people admit this new repartition of tasks.

125. The extent to which the operators can actually complete their job is a real issue, as it depends on their level of accessibility of data from different sources (past data, other surveys, register data, administrative data), and it often happens in practice that they only have a partial access to existing data.

126. Another issue is the quality of manual modifications. Many values are modified, but are we always sure that the modified value is better than the true value? We are not. Some authors underline that it can even be worse (see for example Granquist & Kovar 1997). Why? This job requires training (accounting, classifications), to ensure the quality of the confirmations, codings and updatings. It should also require knowledge about the survey and its objectives, statistics, or quality in general. The validity of manual modifications also depend on the contributors, that might give a plausible but incorrect information, for example to avoid further questioning.

127. Even if we suppose that survey clerks are very well trained, that they understand perfectly well the data they have to verify, that they have access to all kinds of data, it is not obvious that they are able to do their work properly. It happens when they can't get any information from the contributor, and when there are no other available sources of information.

128. A main obstacle to any change in the organisation of manual editing is the common culture of survey operators. It looks the same in many NSIs: the quality is seen as the sum of qualities at the unit's level. Survey clerks try to make sure that every return is as “clean” as possible. If it is not the case, they have the feeling that the overall quality is low. In other words, it often turns out that the distinction between individual quality and statistical quality is not done. This is a real problem because it means that

too much time is spent on each questionnaire, without really taking into account the impact of each unit on the statistics.

129. Last but not least, improving manual editing also requires an improvement of the tools. For example if we want to change the culture, if we want to make people think in a statistical way, it is important to provide them the corresponding information: for example, the data editing software could display, for each unit, its impact in percentage on the target statistics.

## **VI. REDESIGNING THE SURVEY PROCESS**

130. Survey redesign is the last component of a data editing strategy. The idea is to study the results of the data editing process, in order to improve the next survey, as soon as the previous editing process is really completed. Such an analysis highlights mistakes in the way the survey is conducted, and enables us to find where to focus to get a better survey process.

131. The necessity of considering data editing as a tool to improve the process has long been studied in the literature. According to (Granquist & Kovar 1997), the goals of editing are threefold: To provide information about the quality of the data, to provide the basics for the (future) improvement of the survey vehicle, and to tidy up the data. In many papers, L. Granquist emphasises the second aspect (Granquist 1990, 1995, 1997).

### **A. How to “improve the survey vehicle” using information from the last editing process?**

132. One of the first ideas that comes to mind is to use data editing results to tune automatic editing tools: removing some edit rules, relaxing bounds of plausibility intervals, review all kinds of parameters or groupings used for editing.

133. (Lepp & Linacre 1993) mention an improvement which consists in enhancing data editing functions at the beginning, when the respondent is still available, i.e. during data collection. This leads to designing data entry software, using CAPI, CATI or CASIC. However, it is not always suitable for business surveys, particularly for complex ones. A first reason is that data collection has to take account of the heterogeneity of the universe, and hence questionnaires might differ from an industry to another, or from a size band to another, which makes the data entry tools more costly to design. A second reason is that business data might be multi-sourced, and therefore we cannot assume that one interview will complete it.

134. Another idea is to sharpen survey concepts and definitions, and to change the questionnaire, in order to improve response rates and quality of the responses, which has a direct effect on the amount of editing. Redesigning the questionnaire (by changing wordings, order of questions, or by dropping some questions) requires a lot of work upstream, with teams dedicated to questionnaire design and questionnaire testing, like in Canada and Australia for example. To find out where are the main difficulties in the questionnaire, failure rates (for each question) can be used. Another possibility is to classify the different kinds of errors. A typical error that can be found is described in (Hedlin 2001): raw values and edited values are compared on a log scale, and the figure shows very clearly that there are many instances where businesses have responded to the turnover item in actual pounds rather than in £000. Another possibility is to ask subject-matter experts to analyse a subsample of the erroneous answers, in order to understand *why* the contributor did a mistake (what was his reasoning, what was his understanding of the question).

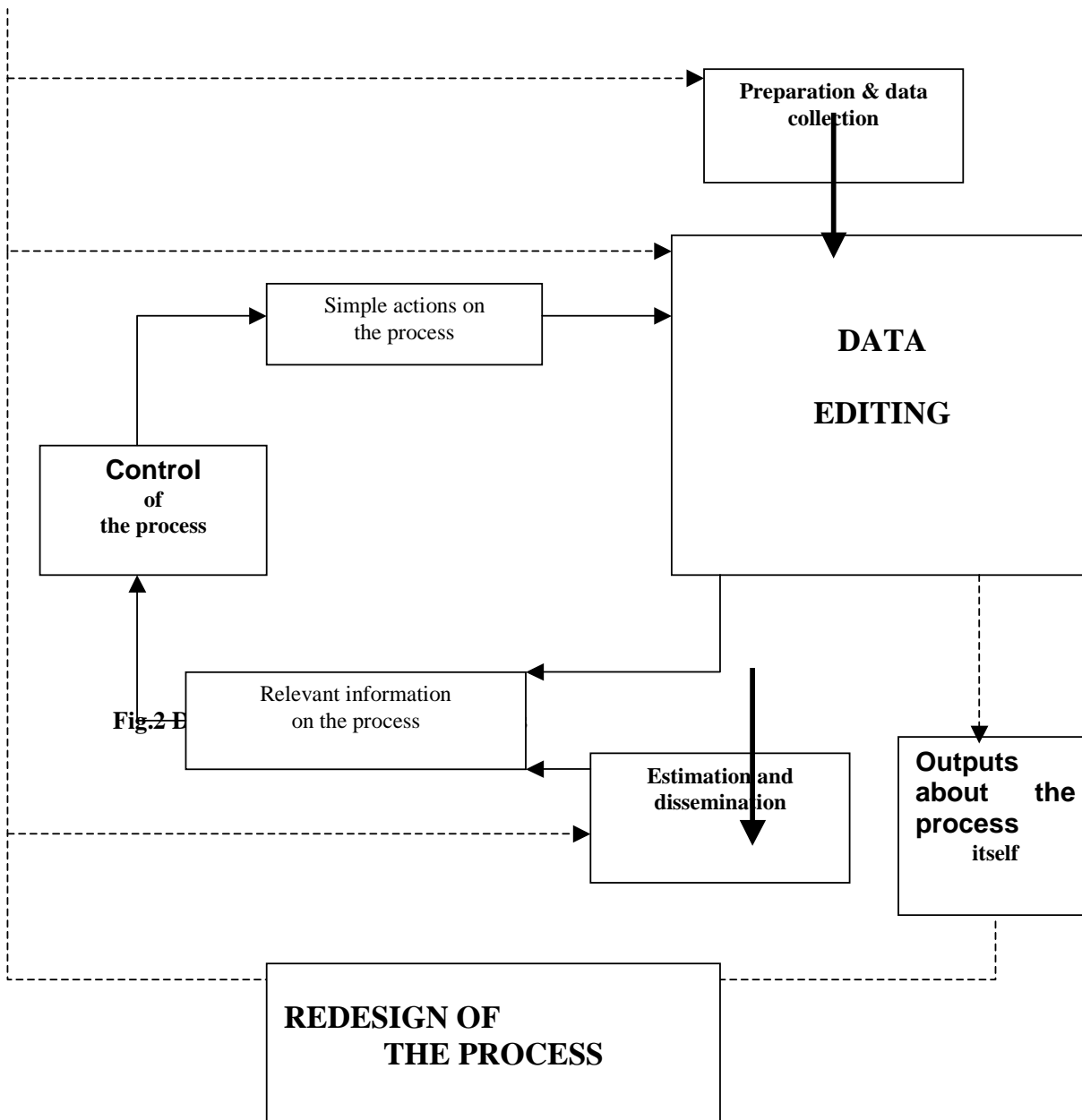
135. Changing the survey process also means better tracking and monitoring, which proves essential in order to give to survey operators a good understanding of the problems encountered and the actions undertaken for a given return. Typically audit trails will consist in keeping track of the changes to values in a field, saving in the same time the reason and source for each change. This gives a way of quickly identifying measurement errors, this knowledge enabling us to make slight changes in the questionnaire,

or event to adapt interviewer training (whenever data collection is done by interview). More generally, adding new metadata allows us to understand how automatic and manual editing actually work.

136. But one mustn't forget the human component of the survey process, as the work done by survey operators can have a major impact on the statistical results (and on the costs, obviously). (Atkinson 2000) stresses the importance of earlier macro-editing capability for survey operators, in first phases of the process, in order to focus as early as possible on records with significant impact. More generally he insists that improving the data editing process means reducing the tendency to compartmentalise: the survey has to be truly integrated, all parts of the system being assembled in a coherent manner, so as to simplify the process and to reduce costs due to lack of communication. Therefore, as pointed out by (Martin 2000), improving the survey vehicle is also retraining survey operators, redefining their tasks, or redesigning the interface of their tools. (Tate & al 2001) insist on the importance of training when moving to a new philosophy of editing, as it helps data editors look positively at new jobs which might require greater skills or flexibility. Downstream, redesigning the survey can also mean redefining users' needs.

137. Applying these ideas in practice obliges one to define quality indicators of data editing (Manzari & Della Rocca 2000), and to use these indicators in a simple and readable manner in order to find out what does not work in the way the information is collected. As an example, failure rates (proportions of doubtful returns) will be used to change the questionnaire, as hit rates (proportion of erroneous returns among the doubtful ones) might prove very helpful to tune edit rules parameters.

**B. Consequence: flexibility of the whole system**



138. The fact that data editing results can provide information to redesign the survey process also has a clear implication on costs: let us stress that *the total cost of editing is not the cost of the data editing process itself* (manual editing being the most expensive part of it). This total cost also comprises the cost of designing the data editing programs and software, the cost of tuning the main parameters, and the cost of redesigning the rest of the process.

139. For example it might happen that some improvements of the process can reduce dramatically the amount of manual editing, but on the same time if these improvements require rewriting the whole application, the real cost is too high. The idea is then to find an intermediate solution, powerful enough but not too complicated to develop.

140. For such a complex process, having the ability of being modified regularly (every survey period) requires that this process is *flexible*. This is a key issue, as stressed by (Esposito & al. 1997), who suggest to “make the system interactive, to increase the flexibility of use”, and “make a system for which it is more natural to make improvements rather than a system for which it is more natural to use and ignore. This means, make a system which gives information that is useful to guiding future improvements.”

141. First of all, the tools have to be flexible enough to be easily tuned. Dropping an edit rule, modifying the parameters of an edit rule, changing the definitions of subpopulations where the parameters are defined, all that has to be simple, which supposes a high degree of generality. But tuning parameters also implies being able to analyse the actual effect of this tuning on the results (failure rate): the automatic data editing process has to be used in a testing environment, for the purpose of simulation, in order to compare estimates using tuned edit rules with estimates using current edit rules (Latouche & Berthelot 1992, Granquist 1995). Consequently, the programs have to be general enough to be applied in two different environments (production and testing). It can prove difficult in practice.

142. But the organization also has to be somewhat flexible. The difficulty of organisational flexibility shouldn't be underestimated, and is probably more important than the tools' flexibility. From the point of view of organisation, optimizing data editing process means “rationalizing the allocation of resources”. But survey operators know that this general expression can have very practical consequences on their day-to-day work: new team, new tasks, new objectives, new priorities, or new tools ... It can't be accepted that easily. As pointed out by (Tate & al 2001, Weir 2000), changing the organisation is not straightforward: “the major barrier to implementation of quality improvement in organisation is resistance to change”.

## **VII. OPTIMIZING DATA EDITING: HOW TO HANDLE THE TRADE-OFF QUALITY/COST?**

143. In the previous sections, we have tried to define what a data editing strategy consists of, elaborating on its five components: principles for selection of doubtful units, prioritization criterion, stopping criterion, manual editing organisation, and redesign of the survey process.

144. In the following we will try to sum up the ways of “optimizing” a data editing process, which means “Having the best possible control of the trade-off between cost and quality”. For that purpose, we will use the data editing strategy as a framework, and for each way of optimizing editing, we will try to analyse to what extent it is possible to put it in practice.

### **A. What do “cost” and “quality” mean**

145. The cost of editing itself can be defined as the number of manually reviewed returns times the average time spent per return (which turns out to be very difficult to measure, and which varies a lot from a survey to another). The total cost should also include: application design, application redesign, parameter tuning.

146. Quality in statistics has many components (relevance, accuracy, timeliness, coherence, comparability, accessibility), but in the case of data editing it is mainly about accuracy. Editing also impacts timeliness, but this is more or less taken into account by the cost dimension. Coherence through time should also be handled, as it is very important for subject-matter statisticians to make sure that changes in aggregates are somehow justifiable, which means that they mustn't be extremely large.

147. Having said that, defining a general indicator for accuracy is not an easy task. A way of doing it consists in defining a list of target variables and domains, a target variation coefficient for each, and then to measure all the corresponding coefficients of variation. These cvs have to take account of sampling error and non-response error, considering all doubtful values as missing values. We can then obtain a list of "standardised" coefficients of variation ( $cv / \text{target } cv$ ). It is clear that some of the standardised variation coefficients may be very high, and sometimes not very relevant, because of many particular cases (restructurings).

148. We can therefore suggest the following global accuracy indicator:

**$Q = 95^{\text{th}}$  centile of the distribution of standardised variation coefficients, over all target variables and all target domains.**

149. It is also important to say that targets in terms of quality and costs are constantly redefined. For example if survey managers find out that there are lots of measurement errors (or a high proportion of item non-responses) on a given variable, it can happen that this variable won't be a target variable any more, which means that it won't take part in the definition of quality. The same goes for costs: there might be sudden budget restrictions, or a political pressure to get results quickly, or changes in the organisation which result in an increase (or a reduction) in resources. Ideally, timeliness should also be included in the definition of quality, as it is essential for the users.

## **B. Optimizing the selection criterion**

150. The general idea is that we want to be able to reduce the number of returns to be checked manually (and therefore reduce the cost of the production process), without increasing significantly the proportion of errors in the end of the process. This means that we want to reduce substantially the failure rate  $f$  and to increase the hit rate  $h$ . Optimizing the selection of doubtful units is a very straightforward way of reducing editing costs. Moreover, it has no influence on the work of survey operators (except that they will have less work), and therefore it does not raise any organisational issue. But how to optimize?

### **Optimizing the edit rules**

151. Optimizing the edit rules practically means optimizing the set of edits (removing useless ones, adding relevant ones) and optimizing the "gates" (bounds of the plausibility intervals) for each edit. An overall optimization of the gates would be a tedious work, as it would have to be done for every edit rule and for all corresponding subpopulations. In practice, it is better to work on a small number of "influential" edit rules, whose failure rate is high (which means that the failure rate of each edit rule has to be saved somewhere). Then after analysing edit rules for which data fail too often, the idea is to measure the impact of changes in the edit rules: what happens if we remove the most influential edit rules? What happens if the gates are enlarged?

152. To perform this kind of optimization, it is crucial to have adequate tools:

- tasks like adding or removing an edit rule, tuning an edit rule (change categories or gates) have to be part of the data editing tool
- the simulation of the data editing process also has to be a feature of the data editing tool: it means that it must be possible to sketch the actual editing process (and therefore to test various sets of edit rules, and various sets of parameters)

- as a result of the simulations and of the data editing process, quality indicators like failure rate and hit rate (for each edit rule) and error rate (whenever the “truth” is known) have to be calculated. These indicators will enable us to determine whether a set of edit rules A is “better” than a set B or not, and are therefore essential if we want to tune the process.

153. These features happen not to be easy to put in practice, as the data editing process is in a production environment, and the design and simulation of the data editing take place in a test environment. In many cases it turns out to be impossible: for example sketching the actual editing process necessitates that the edit rules are clearly written somewhere and really applied, which is not always the case. Having the ability of being tuned means that the tools have to be far more generic and flexible than they usually are.

### **Optimizing the combination of edit rules**

154. As it was said in section 2, the problem is not to optimize the edit rules, but to optimize the global decision rule, at the return’s level. The edit rules are important just because their combination leads to a global rule. We saw that this global rule could be obtained without defining any edit rule (for example with graphical editing).

155. If it is deduced from edit rules, one of the ways of optimizing the selection criterion is to find a good combination of all the edits which is less “defensive” than the approach consisting in taking the intersection of all edit’s acceptability domains. For example if for a given variable, there are 3 minor edits and one major one, and if a unit fails only one minor edit, the variable could be considered as acceptable.

156. From a practical point of view, the same comments can be made as for the optimization of the set of edits: the data editing tool has to be general and flexible, with the possibility to simulate and obtain quality indicators. The tool has to be even a little bit more general and flexible if we want to find an efficient combination of the edits. It can be a lot of additional work: it requires statisticians to determine scores for each edit rule, to determine weights for combining the scores, and maintaining all that. Moreover, if the subject-matter statistician wants to tune the scores, he has to have access to its definition.

### **Imputing instead of editing manually**

157. If a value is considered as erroneous, there are three possible answers:

- doing nothing (which implicitly means validating it)
- checking it manually (and then, either confirming it or updating it)
- imputing the value

158. As explained in section II.D, imputing is a way of reducing the number of returns to be reviewed manually. The underlying idea is that if there are few errors and if their potential impact is small, then imputing instead of verifying would save money without having too large adverse effect on quality.

159. The imputation can be directly related to editing, as in Fellegi-Holt methodology, or not. But whatever the method, imputation has to be used carefully, as “responses to items of interest often have highly skewed distributions, which means that a small number of units contribute substantial amounts to the total estimate” (Granquist 1995). Hence robust imputation methods have to be preferred, and it can prove efficient to measure a distance between the imputed value and the initial value.

160. From an IT point of view, it is not that obvious, because imputation at that stage of the process (on a single unit, just after editing it) is very different from imputation at the end of the process (in order to handle item non-response and total non-response). Imputation at the data editing stage is based on a small amount of information (as the final database is not available), which means the methods used might be different from those used on the final database. The robustness is very important because of the skewness of distributions.

### **Optimizing the peculiar aspects of selection of doubtful units**

161. For the purpose of data editing, a return can be seen as a list of variables, each variable being an answer to a question. But there are some very particular “questions” in a questionnaire. For example, in some surveys, there is a “comments box” in the questionnaire (often at the end of the form), that the contributor fills in whenever he has something to say; survey clerks are generally told to check returns in which this box is filled-in. This means, by definition of doubtfulness, that all returns with filled-in comments box are doubtful. It is clear that such a decision can have a high impact on the amount of data editing, as in some cases, the doubtfulness is only due to the presence of comments. A simple way of eliminating this problem is to get rid of the comments box, or to remove it only for the small units.

162. The same kind of difficulty can happen with some wordings: if there are some noncoded wordings (for example industry wordings), survey operators are told to code them. It also implicitly means that the return is doubtful, and this can substantially increase the failure rate.

163. The underlying idea is that if we want to minimise the cost of editing, it proves very useful to find out whether there are some hidden and obscure reasons why a return might be checked manually even if it seems intuitively acceptable.

### **C. Optimizing prioritization and stopping**

164. In business surveys, the aim is not usually to have a “clean” database, as it would be too expensive. It is important to stress that the real aim of statisticians is to make sure that the statistical results are as accurate as possible, but not to have a “good” database. This implies that the presence of small errors is not a problem. Then in the data editing process, the skewness of the distributions can be seen as an advantage and not as a drawback. The natural idea is to give the highest priority to the most influential units. Such an idea leads to prioritization techniques, units being sorted according to a score that takes into account the impact on statistical results. Obviously, it can be used in order to decide when to stop editing.

165. But the measures of impact can also be seen as a way of defining a selection criterion: the acceptable units will be units whose influence is below a certain threshold. The main difference between such a selection and a stopping criterion based on measures of influence is that the stopping criterion is applied *a posteriori*, as the selection criterion is a way of eliminating units, *a priori*.

### **Impact on aggregates and macroediting**

166. To do that, we have to calculate a score function, taking into account the differences between initial values and plausible values and the impact of those differences on aggregates. A natural stopping criterion can then be defined based on the potential aggregated impact of the remaining (non-checked) units. The problem is that there can be lots of variables, and a breakdown of the population into many domains, which can lead to a huge number of measures of impact. The important point here is to make sure that the information on the current situation of the data editing process is available and readable (using graphical tools for example), in order to take good decisions.

167. Let us stress that the impact on aggregates can be used for prioritization, for stopping, but also for selection: the acceptable units are those whose relative impact is lower than a given threshold. Such an editing technique is a macroediting technique. To do that, it is possible for example to impute all the non-responses and doubtful values, in order to have an early rough approximation, and then to analyse the aggregates (or changes in aggregates, or components of the change) at a high SIC level (for example 2 digits). Then, drilling down in order to find the most erroneous industries at a fine level, and the most erroneous units in it, we can derive early some very influential units. Such a top-down approach is very efficient if combined with graphical tools (Esposito & al. 1997, Weir & al. 1997). It can be seen as another way of performing the selection of doubtful units, without necessarily having to write edit rules, the selection of the non-acceptable units being done graphically. However, this kind of editing has to be

carried out by subject-matter analysts (Granquist 1995), and hence can not be done by traditional survey operators.

### **Impact on mean squared error**

168. Measuring the impact on aggregates and using it to prioritize or to select units for manual checking is clearly useful. But what about the global objective of such a method and its link with “quality”? It is possible to do better: instead of measuring the impact of a difference on aggregates, we can measure for example the potential impact of a given unit on anticipated variance. This requires measuring the variance due to sampling and non-response imputation, and then the potential gain of checking manually a unit. With such an approach, at every moment, the survey manager knows, for each target aggregate, its current value and its current variation coefficient. Moreover, a global quality indicator can be derived from that (see 7.a), and it is then possible *see* the impact of an editing task on this global quality measure. The questionnaires can then be prioritized according to a score, which will be the anticipated gain in terms of variance.

169. Of course this reasoning is not sufficient: we also have to take account of the bias. If bias was not taken into consideration, imputation would generally replace manual editing, as the variance due to imputation is often not very high. Why is it so important to measure the impact on bias? Because some very large values might be correct: therefore if we impute, it can have a high impact on the bias.

170. There are two kinds of problems which have an impact on MSE, and two corresponding actions: doubtful values (→ questionnaire editing) and total non-response (→ non-respondent follow-up). The second action is far more costly than the first one, and is always more efficient in terms of variance; and if we assume that non-response is ignorable, it does not change the bias. Taking into account simultaneously questionnaire editing and nonrespondent follow-up, and knowing the relative costs of each task allow to take good decisions, by measuring, for any given action, the potential impact on mean squared error (or variance) and comparing it to the anticipated cost.

171. Measuring the impact on MSE results into a score that will be very close to the previous “measures of impact on aggregates”. Therefore it can be used for prioritization, as a stopping criterion or for the selection of the doubtful units.

172. But it will have a better theoretical basis as it will be directly linked to quality, which also makes the criterion more readable. Moreover, such an approach, which generalises the previous one, is more complete and precise (use of target cv, distinction between bias and variance, distinction between non-respondent follow-up and questionnaire editing).

173. On the other hand, the technical problems will be a little bit more complicated, as it will be necessary to be able to impute doubtful values and to estimate the variance. In the case of the previous approach (measures of impact on aggregates), as we just want to have an order of magnitude so as to rank the units by decreasing influence, it was even possible to reuse the aggregates obtained at the previous period, which could make things far more simple.

### **Late editing based on changes in aggregates**

174. When analysing the data editing process: we generally think that when the file is “cleaned up”, then aggregates are estimated, and then statistical results are published. That is not true: generally, a huge part of manual data editing comes from the necessity to make sure that the gross flows are not absurd. These changes in aggregates are very important for the users: they do not care about aggregates, they care about changes. If the changes are abnormal, survey operators try to find what is the unit which is responsible for the error. It is implicitly a top-down macroediting approach. As pointed out in section 2.d, “late editing” does not occur late in the process if a graphical macroediting tool is used, as such a tool incorporates the statistical validation. But it requires better skills from survey operators.

175. Late editing means that the criterion for doubtfulness of a return is in a way redefined: units that were acceptable according to the initial selection criterion can be considered as doubtful in this late stage. This also means that manual editing is different at that stage, and requires a better background, for example some statistical knowledge. Therefore, survey operators have to be trained for two operations (current manual editing and late editing) that are basically different, apart from the case of graphical macroediting.

176. The optimization of a data editing system has to take into consideration this kind of editing and all the issues it raises, because such a statistical verification is always done, irrespective of the other techniques used. Moreover, this late validation can turn out to be very costly. This means that the attempt to improve data editing can prove inefficient if the optimization does not take into account in the same time an improvement of the late editing process.

177. And this editing is not an easy one. Changes in aggregates can be difficult to interpret, as they involve: sample rotations, imputations, weight changes, SIC or size changes, births, deaths and other population changes, mergers and demergers, ... and "normal" units (non imputed, remaining in the sample, keeping the same characteristics, etc.). Consequently, the analysis of these changes requires to split them up into their main components if we want to detect potential problems.

178. The importance of being able to interpret changes in aggregates highlights the fact that as far as data editing is concerned "quality" is more than "accuracy": comparability and coherence through time is clearly a component of quality for survey editors. Even if the relative MSE is small, additional editing will be performed whenever the relative change in aggregates is very large. For this reason, a general approach based on impact on MSE will never be sufficient.

#### **D. Reducing the "level of demand" to reduce the amount of editing**

179. Irrespective of organisation, skills and IT, the cost of editing highly depends on what the users want. The amount of manual editing is an increasing function of the number of target domains, the number of target variables, and of the desired quality level. For example, in large surveys, that are used by many different users in many ways, on very small subpopulations, data editing will inevitably prove very expensive.

180. For example, as mentioned in section 4, it is possible to find an approximation of the minimum amount of returns to review manually, as a function of the number of target domains and a kind of quality level. It is easy to see that if the smaller the domain, the larger the proportion of manual checking.

181. Being able to assess the anticipated amount of data editing (even if it is a lower bound, or a very rough approximation) would be very helpful, as it would make things clear for the users. Associating a cost with a level of demand enables one to say something like: if you want to get information at the county level, the cost is x, if you want it for the whole country, the cost is y, y being far smaller than x.

182. The main issue is then to get information on the true demands of the users, to reveal their preferences. The consequence is that in the trade-off between quality and cost, quality can be negotiated. Statisticians shouldn't forget that.

### **VIII. SOME CONCLUSIONS**

183. For a layman, the data editing process looks extremely simple, and hence making it less costly seems straightforward. But a deeper analysis shows that it is a complex operation, and that a thorough understanding of the current process is necessary before trying to improve it.

184. We showed that there were five main components in this process. A way of designing a data editing strategy is to work on each of these components. The details of such a strategy might differ from a

survey to another, as it will depend on the size of the survey, its regularity, its objectives and delays, the level of survey operators, and other aspects.

185. The organizational problems play a key role when putting into practice general ideas on data editing. Who are the survey operators who will review the returns? What is their background? What is their knowledge about statistics, information technology, or about the objectives of the survey? The choice of the approach will depend on the answers to these questions. For example if questionnaires are checked by subject-matter statisticians, or close to that level, output editing is an excellent approach.

186. It should be stressed that graphical editing techniques are particularly efficient, which is not surprising as they have an impact on the five components of the strategy, hence handling several issues in the same time. These techniques enable us to edit and to visualise the outputs in the same time, which means that the trade-off quality/cost is far more visible. Being closer to the outputs and being able to anticipate the impact of the actions is essential if we want to have the best possible control of the process.

187. In practice, it turns out that the development of powerful data editing systems can be difficult, particularly in large and complex surveys. One of the main features that are required is flexibility, because improvements cannot be controlled without modifications of the software and tests of these modifications. This means the ability to simulate the process, and to tune the parameters or the set of edit rules.

188. General software exist (GEIS, CHERRYPI, STEPS, SPEER, etc.), but they are not the ultimate solution: whatever the software, it has to fit with the actual organization, and has to be connected with the existing databases, metadata, applications, habits, which is not always possible. (Pierzchala 1995) proposes a methodical way of evaluating an editing system, with a list of functions and features of such a system, that can be applied either to existing tools or to the development of new tools.

189. Irrespective of the methodology, the tools, or the organization, it is important to have in mind that the starting point of a data editing strategy is the definition of the objectives of the surveys, that have to be negotiated with the users. Quality is fitness for use, it is not a pure concept. Therefore, selection of doubtful units, prioritization, stopping criterion can be defined only according to these objectives. Hence methodologists should be able to link systematically the data editing strategy with the aims of the survey. And if there are many different objectives at a fine-grained level, it should be admitted that data editing is inevitably expensive, very expensive.

## References

- Atkinson D. (2000), Developing a state-of-the-art editing and imputation system for NASS' agricultural censuses and sample surveys, *UN/ECE Work Session on Data Editing*, Working Paper No 27, Cardiff, UK, 18-20 October 2000 (<http://www.unece.org/stats/documents/2000.10.sde.htm>).
- Boydens I., Pirotte A., Zimanyi E. (1997), Managing Constraints Violations in Administrative Information Systems. In *Proceedings of the 7th IFIP 2.6 Conference on Data Semantics, DS-7, Leysin, october 1997*. Chapman & Hall, 1997, p. 241-264.
- Boydens I. (1999), *Informatique, normes et temps*. Bruxelles : Éditions E. Bruylant.
- Chernikova, N.V. (1965), Algorithm for finding a general formula for the non-negative solutions for a system of linear inequalities. *USSR Computational Mathematics and Mathematical Physics*, 5, 228-233.
- De Waal T. (1997), A recipe for applying Cherrypi in the edit process, *UN/ECE Work Session on Data Editing*, Working Paper No 32, Prague, Czech Republic, 14-17 October 1997 (<http://www.unece.org/stats/documents/1997.10.sde.htm>).

- De Waal T., Renssen R., Van de Pol F. (2000), Graphical macro-editing: possibilities and pitfalls, *Proceedings of the Second International Conference on Establishment Surveys*, 17-21 June 2000, Buffalo, New York.
- Engström P., Ängsved C. (1995), A description of a Graphical Macro Editing Application, *UN/ECE Work Session on Statistical Data Editing*, Working Paper No 14, Athens, November 6-9, 1995.
- Esposito R., Lin D., Tidemann K. (1997), The graphical and query system ARIES, in *Statistical Data Editing, Vol. 2, Methods and Techniques*, United Nations, New York and Geneva
- Fellegi I.P., Holt D. (1976), A systematic approach to automatic edit and imputation, *Journal of the American Statistical Association*, 71, 17-35.
- Granquist L. (1990), A review of some macro editing methods for rationalising the editing of survey data, *Proceedings of Statistics Canada Symposium 90*, pp. 225-234.
- Granquist, L. (1995). Improving the Traditional Editing Process. In *Business Survey Methods*, eds. B. Cox, D. Binder, N. Chinappa, A. Christianson, M. Colledge, and P. Kott, New York: Wiley, 385-401.
- Granquist L. (1997), The New View on Editing, *International Statistical Review*, 65, pp. 381-387.
- Granquist L., Kovar J.G. (1997), Editing of Survey Data: How much is enough? In *Survey Management and Process Quality*, New York: Wiley, pp.415-435.
- Hedlin, D. (2001), Selective Editing for Business Surveys. *Technical Report 20/4/01*, University of Southampton.
- Hidiroglou, M.A., Berthelot J.M. (1986), Statistical Editing and Imputation for Periodic Business Surveys, *Survey Methodology*, 12, pp. 73-84.
- Latouche, M., Berthelot J.M. (1992), Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys, *Journal of Official Statistics*, Vol. 8, pp. 389-440.
- Lawrence, D., McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings. *Journal of Official Statistics*, Vol. 10, pp. 437-447.
- Lawrence D., McKenzie R (2000), The General Application of Significance Editing, *Journal of Official Statistics*, Vol. 16, No 3, September.
- Lepp, H., and S. Linacre (1993), Improving the Efficiency and Effectiveness of Editing in a Statistical Agency, *Bulletin of the International Statistical Institute: Proceedings of the 49th Session*, Florence, Italy, Contributed Papers Book 2, pp. 111-112.
- Linacre, S.J. Trewin, D.J. (1993). Total Survey Design – Application to a Collection of the Construction Industry. *Journal of Official Statistics*, 9, 611-621.
- Little, R.J.A., Smith, P.J. (1987), Editing and imputation for quantitative survey data, *Journal of the American Statistical Association*, 82, pp. 58-68.
- Manzari A., Della Rocca G. (1999), A generalised system based on a simulation approach to test the quality of editing and imputation procedures, *UN/ECE Work Session on Data Editing*, Working Paper No 13, Roma, Italy, 2-4 June 1999 (<http://www.unece.org/stats/documents/1999.06.sde.htm>).

Martin C. (2000), Metadata - An aid to managing the edit and imputation process, *UN/ECE Work Session on Data Editing*, Working Paper No 6, Cardiff, UK, 18-20 October 2000 (<http://www.unece.org/stats/documents/2000.10.sde.htm>).

Naus J.I., Johnson T.G., Montalvo R. (1972), A probabilistic model for identifying errors in data editing, *Journal of the American Statistical Association*, 67, pp. 943-950.

Pierzchala M. (1995), Editing systems and software, In *Business Survey Methods*, eds. B. Cox, D. Binder, N. Chinappa, A. Christianson, M. Colledge, and P. Kott, New York: Wiley, 425-441.

Rivière P. (1997) The new annual enterprise surveys in France, *Courrier des statistiques, English Series*, n°3, 1997 annual issue, INSEE, France.

Rivière P. (2001a) What makes business statistics special? *International Statistical Review* (approved for publication).

Rivière P. (2001b) Stopping criterion: a way of optimizing manual editing and assessing its minimal cost, *Technical Report 25/08/01, University of Southampton*.

Tate P., Underwood C., Thomas P., Small C. (2001), Challenges in Developing and Implementing New Data Editing Methods for Business Surveys, *Proceedings of Statistics Canada Symposium 2001*, 17-19 October 2001, Ottawa.

Thompson, K.J., and Sigman, R.S. (1999). Statistical Methods for Developing Ratio Edit Tolerances for Economic Data. *Journal of Official Statistics*, Vol. 15, pp. 517-535.

Underwood C., (2001), Implementing Selective Editing in a Monthly Business Survey, *paper presented at the GSS(M) Conference*, Ambassadors Hotel, London, 25/6/2001.

Underwood C., Small C., Thomas P. (2001), Improving the Efficiency of Data Validation and Editing Activities for Business Surveys, *paper presented at the GSS(M) Conference*, Ambassadors Hotel, London, 25/6/2001.

United Nations (1997), Glossary of terms used in Data Editing, in *Statistical Data Editing, Vol.2*, United Nations, Economic Commission for Europe.

Weir P., Emery R., Walker J. (1997), The Graphical Editing Analysis Query System, in *Statistical Data Editing, Vol. 2, Methods and Techniques*, United Nations, New York and Geneva.

Weir P. (2000), The main barrier to implementing an EDA approach to data editing, *Proceedings of the Second International Conference on Establishment Surveys*, 17-21 June 2000, Buffalo, New York.